

Estimation of Seismic Building Structural Types Using Multi-sensor Remote Sensing and Machine Learning Techniques

Christian Geiß ^{a, b, *}, Patrick Aravena Pelizari ^a, Mattia Marconcini ^a, Wayan Sengara ^c, Mark Edwards ^d, Tobia Lakes ^b, Hannes Taubenböck ^a

^a German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Münchner Straße 20, 82234 Oberpfaffenhofen-Weßling, Germany

tel.: +49 8153 28 1255, fax: +49 8153 28 1445

christian.geiss@dlr.de; patrick.aravenapelizari@dlr.de; mattia.marconcini@dlr.de; hannes.taubenboeck@dlr.de

^b Humboldt Universität zu Berlin, Geography Department, Rudower Chaussee 16, 12489 Berlin, Germany

tobia.lakes@geo.hu-berlin.de

^c Institut Teknologi Bandung, X-PAU Building, Ganesha 10, Bandung, Indonesia

iws@geotech.pauir.itb.ac.id

^d Geoscience Australia, GPO Box 378 Canberra ACT 2601, Australia

mark.edwards@ga.gov.au

Abstract

Detailed information about seismic building structural types (SBSTs) is crucial for accurate earthquake vulnerability and risk modeling as it reflects the main load-bearing structures of buildings and, thus, the behavior under seismic load. However, for numerous urban areas in earthquake prone regions this information is mostly outdated, unavailable, or simply not existent. To this purpose, we present an effective approach to estimate SBSTs by combining scarce *in situ* observations, multi-sensor remote sensing data and machine learning techniques. In particular, an approach is introduced, which deploys a sequential procedure comprising five main steps, namely calculation of features from remote sensing data, feature selection, outlier detection, generation of synthetic samples, and supervised classification under consideration of both Support Vector Machines and Random Forests. Experimental results obtained for a representative study area, including large parts of the city of Padang (Indonesia), assess the capabilities of the presented approach and confirm its great potential for a reliable area-wide estimation of SBSTs and an effective earthquake loss modeling based on remote sensing, which should be further explored in future research activities.

Keywords

Seismic Building Structural Types, very high and medium resolution imagery, machine learning, SVM, Random Forests, earthquake loss estimation

1. Introduction

Increasing spatial concentration of exposed elements such as people, buildings, infrastructure or economic values in earthquake prone regions induce seismic risk at an unprecedented high level. In particular, urban areas in developing countries are characterized by a large amount of vulnerable buildings. At the same time, a very dynamic urban growth is accompanied by the construction of unplanned, spontaneous and highly vulnerable settlements. Thus, local governments and stakeholders face the problem of continuously updating their knowledge on the current building stock and simultaneously assessing exposed buildings area-wide to efficiently establish and adjust preparedness measures (Sarabandi and Kiremidjian, 2007, Taubenböck et al., 2009a, Wieland et al., 2012). Especially for earthquake loss estimation (ELE) modeling, the gathering of building inventory and vulnerability information represents normally the most time-consuming and expensive aspect (Dunbar et al., 2003).

The exclusive application of conventional approaches such as detailed *in situ* building-by-building analysis by structural engineers is decreasingly able to cope with this situation. Instead, in the last few years remote sensing has proven its great potential to extract relevant features for pre-event vulnerability analysis of built-up structures for large areas (Geiß and Taubenböck, 2012). So far, different approaches have been presented in the literature. By means of characteristics extracted from remote sensing data, Taubenböck et al. (2009a) and Borzi et al. (2011) reconstruct and characterize the built environment and retrieve specific fragility functions for designated building types. Pittore and Wieland (2012) use remote sensing data for delineating and characterizing homogeneous built-up areas. The vulnerability of the building inventory is determined in combination with information from a ground-based omnidirectional imaging system. Similarly, Borfecchia et al. (2009) assess the vulnerability of buildings in a hybrid way, namely by combining *in situ* ground truth for selected buildings with information derived from remote sensing data. Supervised classification techniques are subsequently used to classify the residual building inventory. Geiß et al. (2013) combine detailed *in situ* seismic vulnerability information with features describing the urban morphology derived from remote sensing data. Supervised regression and classification techniques are then applied to evaluate the suitability for an area-wide assessment. The aforementioned studies deploy very heterogeneous approaches with respect to the vulnerability levels or classes to be estimated. Taubenböck et al. (2009a), Borzi et al. (2011), and Borfecchia et al. (2009) use rather specific definitions, whereas Pittore and Wieland (2012) and Geiß et al. (2013) incorporate also more transferable, yet generalized, assessment schemes, such as the European Macroseismic Scale (EMS-98; Grünthal et al. 1998). However, none of the cited studies focus on the estimation of seismic building structural types (SBSTs). SBSTs characterize the main load-bearing structure of a building. This is the most affecting factor for earthquake damage and, accordingly, it is generally the first property considered for categorizing a building. Further frequently considered parameters that may reflect the seismic performance also comprise the number of storeys, the period of construction or the presence of structural irregularities (Coburn and Spence, 2002). A function for individual SBSTs can be determined that relates the magnitude of the seismic hazard to the damage probability of the structures (Calvi et al., 2006). This enables the prediction of the

probable damage distribution of the building inventory with respect to a certain level of seismic hazard (Douglas, 2007). Additionally, SBSTs can also contribute to the assessment of the seismic vulnerability according to schemes such as the EMS-98.

In the pioneering work of Sarabandi and Kiremidjian (2007), information derived from remote sensing data is combined with ancillary (geo-)information to estimate SBSTs. In particular, they use very high resolution optical imagery to derive the building inventory and calculate features describing the height, extent, shape, and roof type characteristics of individual buildings. In addition, they use tax assessor data to compile information about occupancy and age. Subsequent to that, supervised classification techniques (Classification and Regression Trees (CART) and multinomial logistic regression) are deployed to estimate SBSTs. In this paper we propose considerable conceptual and methodological differences to estimate SBSTs. The plethora of sensors systems that provide useful and complementary information yields the possibility to substitute ancillary (geo-)information and, thus, fully rely on remote sensing to reconstruct and characterize the building inventory. Due to e.g. data availability it may be crucial to gain independence from proprietary sources of information (e.g. tax assessor data). In addition, a complementary set of remote sensing data allows to characterize the building inventory in an exhaustive manner and to encode for instance also spatial context information in the classifier. This in turn opens a good opportunity to boost predictive performance of learned models. An exhaustive characterization of the building inventory based on a comprehensive set of features simultaneously suggests relying on classification approaches that are able to cope efficiently with high-dimensional data sets. Moreover, SBSTs ground truth is very costly to obtain and at the same time is afflicted with uncertainties induced by an often challenging assignment process. This induces the general need for a more tailored approach, which is able to lower those uncertainties and can cope with the scarcity of *in situ* elaborated ground truth.

To address these considerations, the objective of this paper is to introduce an approach for estimating SBSTs area-wide based on scarce *in situ* ground truth and complementary multi-sensor remote sensing data by means of a sequential procedure of advanced machine learning techniques. More specifically, we exploit very high resolution multispectral imagery, multi-temporal medium resolution multispectral data, as well as height information from a normalized digital surface model (nDSM) to derive a comprehensive set of features characterizing the urban environment. Different feature selection techniques are then employed to reduce the dimensionality of the resulting dataset and identify the most relevant features. Outlier detection is applied to prune those objects from the data for which the available *in situ* information cannot be considered reliable. To tackle the scarcity of SBSTs ground-truth data, additional synthetic samples are generated. Finally, different SBSTs are estimated by means of advanced supervised classification techniques. In particular, both Support Vector Machines (SVM) (Vapnik, 1998, Schölkopf and Smola, 2002) and Random Forests (RF) (Breiman, 2001) are considered due to their capability of effectively handling complex remote sensing classification problems (Camps-Valls and Bruzzone, 2009, Gislason, 2006). Since spatially distributed estimation of SBSTs represents a critical input for ELE models, we illustrate the applicability of the presented approach within scenario-based loss estimations for the city of Padang, Indonesia.

The remainder of the paper is organized as follows. [Section 2](#) characterizes the study site and data base for this study. In [Section 3](#) the methods are described and results and discussions are revealed in [Section 4](#). The paper is concluded and an outlook is given in [Section 5](#).

2. Study Site and Data

2.1. Study Site: Padang, Indonesia

The presented study focuses on the city of Padang (Indonesia), which is situated in one of the most earthquake-prone regions worldwide. Padang is located on the island of Sumatra (mainly on the coast and to some extent sited beneath the mean sea level) and is the capital city of the Sumatera Barat province. It represents the third largest city on the island with approximately one million inhabitants. The dynamic urban system of Padang is constituted by a high concentration of population, infrastructure and economic values. The city has supra-regional relevance with an international airport and railway connection and possesses an essential economical role for the coastal region and the mountainous hinterland.

The Sumatra subduction zone represents one of the most active plate tectonic margins in the world ([Petersen et al., 2004](#)). The Australian plate plunges beneath the Sunda block of the Eurasian plate with convergence rates between ~ 56 and 62 mm/yr. ([Chlieh et al., 2008](#)) ([Fig.1a](#)). The associated complex plate boundary setting leads to thrust earthquakes on the subduction fault, strike-slip earthquakes on the Sumatran fault, deeper earthquakes within the subducting lithosphere, and volcanic earthquakes ([McCaffrey, 2009](#)). Accordingly, the city is located in a region characterized by extremely high probability of severe earthquakes, as well as secondary effects such as tsunamis ([Chlieh et al., 2008](#), [Taubenböck et al., 2009b](#)). As an example, in the afternoon of 30th September 2009, Padang was hit by an earthquake with a moment magnitude of $M_w = 7.6$ ([Taubenböck et al., 2013](#)). Overall, the earthquake event affected an area with a population of 1.2 million and caused 1,195 fatalities. 144,000 buildings collapsed or were significantly damaged. In Padang, 383 people died and 431 were seriously injured, primarily due to collapsing buildings ([EERI, 2009](#), [BNPB, 2009](#)). Despite the size of the event, the Sunda megathrust was not ruptured and the stress on the Mentawai segment, which was accumulated over 200 years, has not been significantly reduced. The megathrust strain-energy budget remained substantially at a high level and the threat of a great, also tsunamigenic earthquake with a magnitude $M_w > 8.5$ on the Mentawai patch is unabated ([McCloskey et al., 2010](#)).

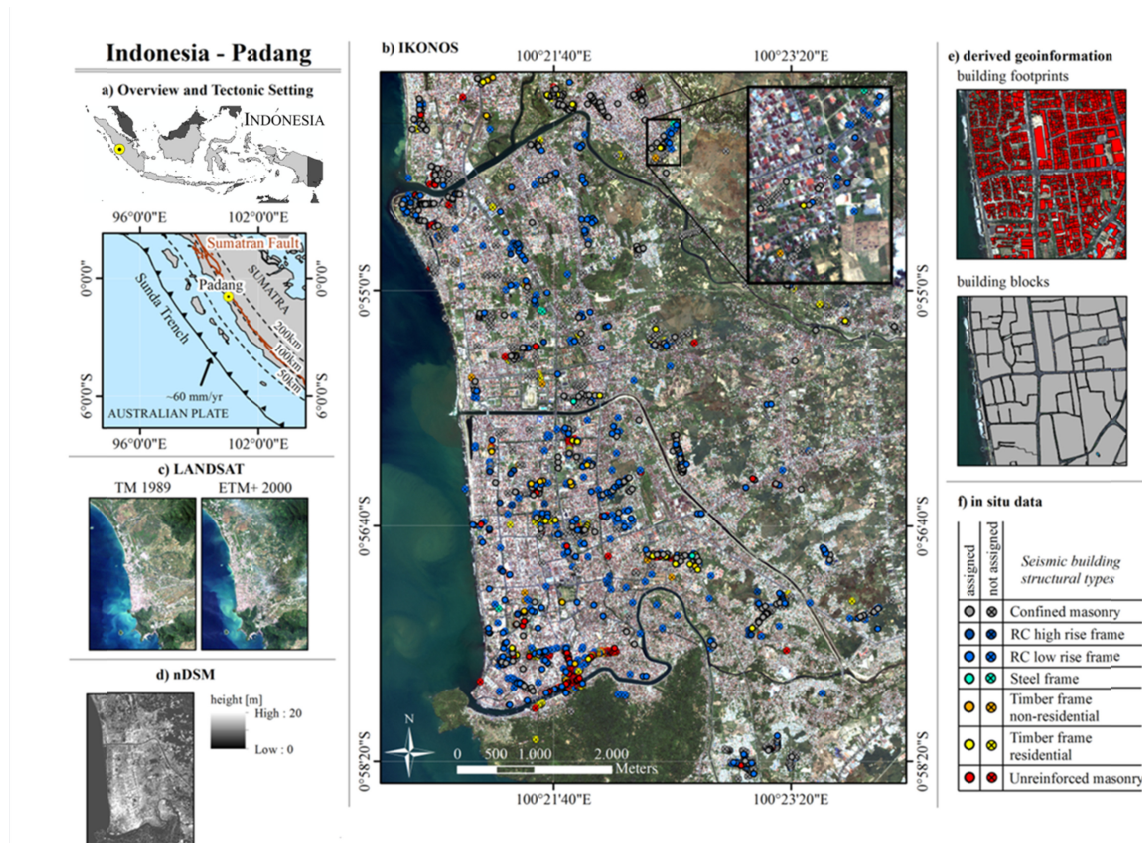


Fig. 1. Overview on the location of the study area and acquired data. (a) overview on the location of Padang and tectonic setting; basic active structural elements of the obliquely convergent Sumatran plate boundary are shown; dashed lines parallel to the trench are the 50, 100, and 200km depth of the megathrust (Chlieh et al., 2008); (b) very high resolution multispectral IKONOS imagery; (c) multitemporal LANDSAT data; (d) nDSM with object heights; (e) derived geoinformation consisting of building footprints and buildings blocks; (f) *in situ* data regarding SBSTs that are superimposed on (b).

2.2. Remote Sensing Data

The remote sensing data used in our study have been acquired and processed within the “Last-Mile” project (Taubenböck et al., 2009b) and comprise a multispectral IKONOS image, multitemporal LANDSAT data, as well as height information from a nDSM (all of them co-registered to UTM 47S projection and WGS-84 datum). The multispectral IKONOS image was acquired on 12th April 2005 and covers a spectral range between 0.445 and 0.853 μm , with a geometric resolution of 1 m for the panchromatic band and 4 m for the 4 multispectral bands. The data were pan-sharpened and atmospheric correction was performed using the ATCOR (Atmospheric and Topographic Correction) model (Richter, 1996, *ibid.*) (Fig. 1b). Multitemporal LANDSAT data were acquired by the Thematic Mapper sensor and the Enhanced Thematic Mapper sensor on 25th July 1989 and 15th July 2000, respectively. Both images are characterized by 7 multispectral bands covering a spectral range between 0.45 and 2.35 μm at 30 m spatial resolution (Fig. 1c). Height information is derived by means of a digital surface model (DSM) and a digital terrain model (DTM), both derived from airborne radar data acquired by pair antennas and processed using SAR interferometry techniques (Li

et al., 2004). The DTM was derived based on measurements of the bare ground contained in the original radar data and by manually reviewing and editing (Intermap, 2010). The data sets have a geometric resolution of 5 m and a height Root-mean-square-error (RMSE) of 1 m. To get relative height information of elevated objects a nDSM is calculated by subtracting the height values of the DTM from the height values of the DSM (Fig.1d). Interested readers can refer to Taubenböck et al. (2009b) for a more detailed description of data acquisition and preprocessing.

2.3 Derived Geoinformation and In Situ Data

Basic geoinformation had already been derived from the remote sensing data and provided for this study. Within the “Last-Mile” project 87,573 building footprints were digitized from the IKONOS imagery by means of a manual photointerpretation procedure. They represent the core of Padang’s building inventory. Additionally, building blocks had been derived from a closed-meshed road network (Taubenböck et al., 2008) (Fig. 1e). Both information layers serve as basis for the calculation of features, which is explained in Section 3.1.

About four weeks after the earthquake on 30th September 2009 a field survey in the affected region took place in the framework of the Australia-Indonesia Facility for Disaster Reduction (AIFDR) jointly led by the Institut Teknologi Bandung (ITB) and Geoscience Australia. The primary objective of the survey was to undertake a population based inspection of buildings of all types and all damage levels. The results allowed inferring knowledge regarding the vulnerability of a range of building types present in the surveyed region and representative for others in Indonesia (Sengara et al., 2010). Overall, 3896 buildings were surveyed and each of them was assigned to a specific structural system, wall type, roofing type, floor type, number of storeys, usage, and the degree of damage suffered from the earthquake event. To conduct a vulnerability assessment and to derive fragility curves, the surveyed buildings were categorized according to SBSTs, which reflect a similar behavior under seismic load. In particular, the following classes were considered: “Confined masonry” (*CM*), “Reinforced concrete high” (*RC high*), “Reinforced concrete low” (*RC low*), “Steel frame” (*SF*), “Timber frame residential” (*TF res*), “Timber frame non-residential” (*TF non-res*), and “Unreinforced masonry” (*URM*) (Fig. 1f). From the whole amount of surveyed buildings, 2779 are located in the study area. The position of each building was recorded with a GPS device (Fig. 1b) and digital pictures were also taken (ibid.). Nevertheless, due to inaccuracies in the GPS positioning, only 561 buildings could be unambiguously assigned to their corresponding building footprint extracted from remote sensing imagery (Fig. 1e). Unfortunately, only two samples remained for the structural type *SF*. As this class represents a relatively rare but striking SBST, the corresponding *in situ* ground truth was extended with 12 additional samples derived from another data set compiled in February/March 2008 during the “Last-Mile” project (Taubenböck et al., 2009b). The histogram depicting frequencies of different SBSTs of the final *in situ* data set is shown in Figure 2. Descriptive statistical analyses were carried out to check whether the final *in situ* data set is consistent with all surveyed buildings and the results revealed a very good agreement.

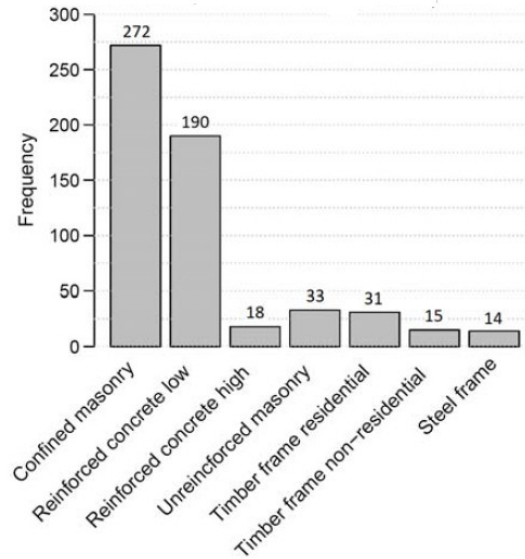


Fig. 2. Frequency of labeled samples (overall: 573) according to different SBSTs of the final *in situ* data set

3. Methods

Based on the remote sensing and *in situ* data, we carry out a sequential procedure to estimate SBSTs. Figure 3 gives a schematic overview from the data sets used, the chronology of the procedure to the targeted SBSTs classification map. A set of features is derived from the remote sensing data at two different spatial levels, building and block level (Section 3.1). The hierarchical supervised classification approach is described in Section 3.2. Outliers in the *in situ* data and building inventory are identified first. Therefore, a subset based feature selection technique (Section 3.2.1) is used to create a suitable group of features for building robust one-class classification models based on the *in situ* data. The models are built by means of a one-class support vector machine (OC-SVM, Section 3.2.2) approach and are applied on both *in situ* data and building inventory. Subsequent to outlier identification, multiclass classification models are built in three consecutive steps. The remaining *in situ* samples are used to identify useful groups of features for building robust models by applying subset and ranker based feature selection techniques (Section 3.2.1). To tackle scarcity of the *in situ* data and learn efficient discriminative classifiers, synthetic training samples are generated by means of an oversampling technique (Section 3.2.3). Based on the generated feature groups and oversampled training data, multiclass classification models are learned by using SVM and RF (Section 3.2.4). Finally, the most accurate model is applied on the building inventory to estimate SBSTs spatially distributed.

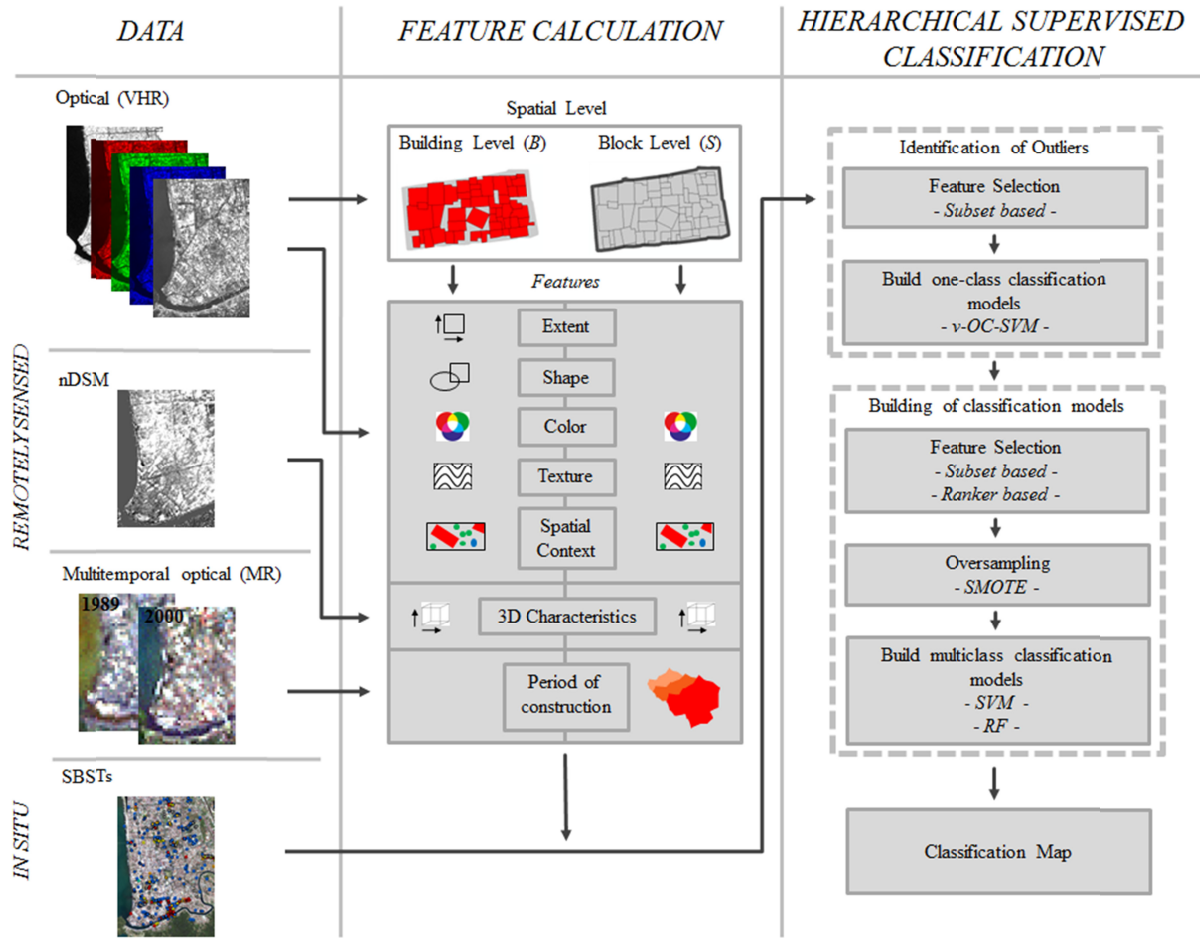


Fig. 3. Overview of the framework and processing steps followed in this study. Detailed explanations are given throughout the text in [Section 3.1](#) which describes the feature calculation, and [Section 3.2](#) which describes the hierarchical supervised classification approach.

3.1. Calculation of Features from Remote Sensing Data

For a reliable estimation of SBSTs, numerous features have been extracted for both the aforementioned individual building footprints and building blocks (identified in the following by the subscripts B and S , respectively) (Fig. 1e & 3). Generally, the building footprints allow characterizing individual buildings, whereas the building block layer characterizes the spatial setting which the respective buildings are embedded in. We chose to use building blocks derived from a street network rather than artificial spatial units, such as quadratic objects. This allows us to reflect the urban morphology, which is constituted by distinct areas that are generally irregularly shaped, more naturally. Simultaneously, the difficulty of having to determine the optimal kernel size *a priori* is avoided (Herold et al., 2003).

We use a set of features, that was introduced and explained in detail in Geiß et al. (2013), where it was used to evaluate the potential of remote sensing to assess the seismic vulnerability levels of buildings (see Fig. 4a). In particular, the features relate to the two-dimensional extent of buildings as well as the description of their shape characteristics. In

addition, statistical values of 1st and 2nd order were extracted from the available IKONOS imagery at building and block level. The first serve as a descriptor of roof surface material and arrangement whereas the latter are intended to describe the composition of distinct urban structures. Mean and standard deviation values of the different image bands as well as band ratios, which are intended to emphasize spectral dissimilarities, were calculated. Additionally, rotation-invariant texture measures for the panchromatic and near-infrared band were computed using both the co-occurrence matrix (GLCM) and grey level difference vector (GLDV). Features explicitly aiming to describe the spatial context are calculated at block level and consist of the area of building blocks and the average size of the buildings located within. Furthermore, spatial metrics such as proportion measures of land cover classes are computed. Based on a urban land cover map derived in Taubenböck et al. (2009b) (which exhibited an OA of 97%), proportions of land cover classes “buildings”, “sealed”, “grass/meadow”, “trees”, and “impervious surface”, which represents a combination of “buildings” and “sealed” were calculated per block. Additionally, a semantic classification (“Structure Type s ”), which is built on physical features that describe the urban morphology, is incorporated. The classification describes the socio-economic status of the population by distinguishing “slums”, “suburbs”, “low income areas”, “medium income areas”, and “high income areas”. Beyond, the incorporation of height information allows the calculation of 3D features such as building floor number, floor space, ratio of diameter and height, ratio of width and height, as well the average building height within a building block. The mean slope for each building block was calculated to describe topographic location characteristics. By analyzing two Landsat images from 1989 and 2000, the period of construction is approximately described based on a post classification change detection procedure, which aims to map the urban extent at the respective time step. For a more comprehensive description of all the features listed in Fig. 4a, the reader is referred to Geiß et al. (2013).

In addition, we introduce a number of new features in this study (Fig. 4b). They characterize the buildings’ two-dimensional extent, shape and spatial context. They were chosen since they turned out to be beneficial in previous studies for discriminating different morphologically homogeneous urban structures based on geospatial data (Steiniger et al., 2008, Colaninno et al., 2011). Overall, each building object is represented by a 145-dimensional feature vector, whereby 79 features are calculated based on the individual building footprints, and 66 are calculated based on the building blocks. In this manner, a perceptual coherence (Steiniger et al., 2008) of physical appearance, spatial composition and context, and temporal development of the urban morphology and the main load-bearing structure of buildings is assumed.

FEATURES

a)	IKONOS Extent/Shape	IKONOS Spectral 1 st order	IKONOS Spectral 2 nd order	IKONOS Spatial context
	Area _B	Mean Max. Diff. _{B, S}	GLCM Angular 2 nd Moment _{B pan nir, S pan nir}	Area Building Block _S
	Perimeter _B	Mean Brightness _{B, S}	GLCM Contrast _{B pan nir, S pan nir}	Average Building Size _S
	Width _B	Mean Blue(1) _{B, S}	GLCM Dissimilarity _{B pan nir, S pan nir}	Degree of Building Density _S
	Length _B	Mean Green(2) _{B, S}	GLCM Std. Dev. _{B pan nir, S pan nir}	Share of LC class "sealed" _S
	Length/Width _B	Mean Red(3) _{B, S}	GLCM Mean _{B pan nir, S pan nir}	Share of LC class "grass/meadow" _S
	Asymmetry _B	Mean Nir(4) _{B, S}	GLCM Homogeneity _{B pan nir, S pan nir}	Share of LC class "tree(s)" _S
	Elliptic Fit _B	StDev. Blue(1) _{B, S}	GLCM Entropy _{B pan nir, S pan nir}	Share of Impervious Surface _S
	Radius of Smallest Enclosing Ellipse _B	StDev. Green(2) _{B, S}	GLCM Correlation _{B pan nir, S pan nir}	Structure Type _S
	Radius of Largest Enclosing Ellipse _B	StDev. Red(3) _{B, S}	GLDV Angular 2 nd Moment _{B pan nir, S pan nir}	
	Rectangular Fit _B	StDev. Nir(4) _{B, S}	GLDV Contrast _{B pan nir, S pan nir}	
	Roundedness _B	[M(4)-M(3)]/[M(4)+M(3)] _{B, S}	GLDV Mean _{B pan nir, S pan nir}	
	Density _B	Brightness×M(1)/M(2) _{B, S}	GLDV Entropy _{B pan nir, S pan nir}	
	Main Direction _B	Brightness×M(1)/M(3) _{B, S}		
	Border Index _B	Brightness×M(2)/M(3) _{B, S}		
	Shape Index _B	M(1)/M(2) _{B, S}	nDSM 3D	LANDSAT temporal
	Compactness _B	M(1)/M(3) _{B, S}	Floor Number _B	Period of construction _S
		M(1)/M(4) _{B, S}	Floor Space _B	
		M(2)/M(3) _{B, S}	Diameter/Height _B	
		M(2)/M(4) _{B, S}	Width/Height _B	
		M(3)/M(4) _{B, S}	Average Building Height _S	
		M(2)/[M(1)+M(2)+M(3)+M(4)] _{B, S}	Slope _S	
		[M(4)/M(3)]/[M(1)+M(2)+M(3)+M(4)] _{B, S}		
		[M(3)-M(2)]/[M(3)+M(2)] _{B, S}		
		[M(3)-M(1)]/[M(3)+M(1)] _{B, S}		
b)	IKONOS Extent/Shape	Description	IKONOS Spatial context	Description
	Area/Perimeter _B	Area to perimeter ratio	Area/Perimeter _S	See description at "IKONOS Extent/Shape";
	Shape Index II _B	Schumm's longest axis to area ratio (MacEachren 1985)	Shape Index II _S	Calculated for Block Level
	Shape Index III _B	Building perimeter to perimeter of a circle with the same area ratio; Gravelius's compactness measure (Gravelius 1914)	Shape Index III _S	
	Building Elongation _B	Length to width ratio of the minimum bounding rectangle (Steiniger et al. 2008)	Building Elongation _S	Homogeneity of alignment
	Building Orientation _B	Orientation of the major axis of the minimum bounding rectangle	Building Orientation _S	
			Proximity Sum _{B, S}	Summed/Averaged distance to 5 nearest neighbors
			Proximity Avg. _{B, S}	
			No. Buildings Buffer 100m _B	Count of all buildings intersected by a buffer with 100m and 200m distance (Steiniger et al. 2008)
			No. Buildings Buffer 200m _B	
			Area to Buffer Area 100m _B	Area to buffer area ratio for a buffer with 100m and 200m distance (Steiniger et al. 2008)
			Area to Buffer Area 200m _B	

Subscripts: B = Building Level, S = Block Level, Pan = Panchromatic, Nir = Near-Infrared; M = Mean Value

Fig. 4. Features derived from remote sensing data to characterize the urban morphology; a) set of features extracted as described in Geiß et al. (2013); b) additional features that are related to the extent and shape of individual buildings and the spatial context they are embedded in.

3.2. Hierarchical Supervised Classification Approach

3.2.1 Feature Selection with Filters

For taking into account the high dimensionality of the input data, several feature selection algorithms were deployed. The selection of features to be used for building robust classification models is generally a difficult task, especially when dealing with a large number

of features as it is the case in this study. Thereby, high-dimensional feature vectors often exhibit redundancy, show inter-correlations and may suffer from the “Hughes phenomenon” (which states that for a limited amount of samples the predictive power decreases as the dimensionality of the feature vector increases (Hughes, 1968)). Furthermore, multivariate classification methods can be prone to over-fitting. Filtering out the least promising features and thus reducing the dimensionality of the feature vector may attenuate the aforementioned problems (Guyon, 2003).

Feature selection techniques can be categorized into filters and wrappers. The latter evaluate features by using accuracy estimates provided by the actual classification algorithm, which is deployed subsequent to feature selection. Thus, the classifier needs to be trained and accuracy estimation needs to be performed for each iteration of the evaluation process. This can lead to large processing times what make them unpractical for extensive studies on high-dimensional data sets (Kohavi and John, 1997). Accordingly, in our study we employed filter methods for feature selection, as they operate independently with respect to the classifier. Among filter methods one can discriminate algorithms which evaluate individual features and those which assess subsets of features (Hall and Holmes, 2003). Overall, six different filter methods were applied in the context of the presented work (by means of the WEKA software environment (Hall et al., 2009)), namely Information Gain (*IG*), Gain Ratio (*GR*), Chi-Squared (χ^2), Pearson product-moment correlation coefficient (*COR*), and Relief-F (*RelF*), which all allow feature ranking, as well as the Correlation-based Feature Selection (*CFS*) which allows evaluating feature groups.

IG belongs to the group of information theory indices, which evaluates a feature by means of the Shannon entropy. The decrease in entropy of a class is evaluated when a feature is considered (Duch, 2006). A drawback of the *IG* measure is that it tends to be biased towards features with a large number of distinct values. *GR*, which is a largely-employed modification of *IG*, aims to avoid this bias by a simple normalization utilizing the class entropy (Van Hulse, 2009). The χ^2 method deploys the χ^2 statistic to evaluate the strength of the relationship between each independent feature and a class (ibid.). *COR* measures the relation of a feature and a class based on Pearson’s Correlation coefficient (Hall et al., 2009) and was used as a benchmark in this study. Focusing on the expectation that useful features should differentiate between instances from different classes and have similar values for instances from the same class, *RelF* (Kononenko, 1994, Robnik-Šikonja and Kononenko, 2003) ranks features according to their ability to discriminate between neighboring instances. For multiclass problems *RelF* randomly samples an instance from the data and locates its k nearest neighbors from the same and different classes. The feature values of the nearest neighbors are compared to the sampled instance and used to up-date relevance scores for each feature. This procedure is repeated for a number of instances m that has to be specified by the user (Hall and Holmes, 2003). The *CFS* method (Hall, 1999) uses a best first search algorithm to identify a group of possibly suitable subsets. Each considered subset is subsequently merited by means of an entropy based heuristic. Subsets with high feature-class correlation and low feature-feature inter-correlation are merited best (Hall and Holmes, 2003). *IG*, *GR*, χ^2 , and *COR* are bivariate procedures. Each feature is evaluated independently of all other features in the data set. This is in contrast to the procedure of the *RelF* and *CFS* approach.

For feature selection methods that require discrete values (i.e. IG , GR , χ^2), numerical features were discretized according to the method of [Fayyad and Irani \(1993\)](#). Concerning the *RelF* approach, we set m to the number of all instances, since a larger value of m implies a more reliable approximation ([Kononenko, 1994](#)). We tested several values for the number of nearest neighbors to be considered. However, the results are hardly sensitive with respect to this parameter in this study, thus the number of neighbors k was set to 10. For the *CFS* method we used a stopping criterion for the search heuristic of 5 consecutive fully expanded non-improving subsets. The subset with the highest merit revealed during the search was selected. The results of the feature selection algorithms served for the creation of feature subsets that were used for building several multiclass classification models ([Section 3.2.4](#)). 14 subsets containing the n “best ranked” features ($n = 5, 10, \dots, 50, 60, 80, 100, 120$) were created based on the results of each ranker method. In addition, one feature subset was built according to the result from the *CFS* technique ($n = 28$). Altogether this leads to a total amount of 72 data sets, one containing the original number of features and 71 represent feature reduced subsets.

3.2.2 Outlier Detection with OC-SVM

Outlier detection is applied to exclude objects from the data for which the available *in situ* information cannot be considered reliable. The appearance of outliers can be related to several sources. The set of *in situ* samples may not be fully representative for all buildings in the area and the SBSTs, as defined in the context of the *in situ* survey, may not cover all structures. Inaccuracies may have been occurred in the surveying of single buildings, leading to the assignment of inappropriate labels. Mislabeling also cannot be excluded in the assignment process of the *in situ* data. Beyond, feature values of building objects can be biased due to errors or noise in the underlying remote sensing data.

To lower those uncertainties we employ ν -OC-SVM, which were introduced by [Schölkopf et al. \(1999\)](#) as support vector method for novelty detection. It is based on the general principle of SVM, which determine appropriate parameters that construct a decision surface, the optimal separating hyperplane, between the classes of training samples with respect to their position in an n -dimensional feature space ([Vapnik, 1995, 1998](#)). With the benefits of SVM, the one class classifier is able to capture the support region (i.e., where the density is large) without the need of prior assumptions about the distribution of the data. Therefore, the target class is described by a function that maps the majority of instances to a region where the function is nonzero. To achieve this, the origin of the feature space is first treated as the only available member of the non-target class (i.e., as an outlier). Then, a hyperplane with maximum margin separation from the origin is identified. Analogous to the usual SVM framework, outliers in the training data are handled by *slack variables* ([Muñoz-Marí et al., 2010](#)).

To keep computational costs low and mitigate over-fitting, we identified a group of valuable features for the one-class models by using the *CFS* technique ([Section 3.2.1](#)). The *in situ* data with the reduced set of features were split according to the SBSTs class labels. Subsequently, for each of them we determine a specific ν -OC-SVM model (all support vector methods in

this paper were carried out with the LibSVM package by [Chang and Lin \(2001\)](#)). We used Gaussian radial basis function (RBF) kernels as commonly used in literature when addressing environmental applications since they showed generally good performances in plenty of studies (e.g. [Camps-Valls and Bruzzone, 2009](#), [Volpi et al., 2013](#)). The application of the ν -OC-SVM approach with a RBF kernel requires adjusting the parameter ν , which represents the expected percentage of outliers in the training set and the kernel-width parameter γ . However, it is difficult to tune free parameters if only target labeled samples are available in the training data. In such situations solely the true positive rate (sensivity) can be calculated, whereas the error counterpart (specificity) cannot. To overcome this limitation, the free parameter selection was determined by evaluating $\arg \max_{\theta} \left\{ \frac{OA[\%]}{\#SV} \right\}$, where θ is the set of free parameters (i.e., ν and γ), OA is the overall accuracy and $\#SV$ the number of support vectors. This heuristic enforces high OA while simultaneously limiting model complexity keeping a low number of SV ([Muñoz-Marí et al., 2010](#)). For both ν and γ we performed a grid search varying ν in the range $\{0.01, \dots, 0.1\}$ in 0.01 steps and γ in the range $\{10^{-2}, \dots, 10^2\}$ in power of $\sqrt{10}$ steps, respectively. OA for each model was estimated by a 4-fold cross-validation strategy. [Table 1](#) shows the determined hyper-parameter combinations for the respective SBSTs with number of SV , OA estimates and corresponding values of the evaluation heuristic.

SBSTs	ν	γ	$\#SV$	OA[%]	OA[%]/ $\#SV$
CM	0.01	0.01	8	96.31	12.04
RC low	0.01	0.01	5	95.74	19.15
RC high	0.01	0.02	5	66.67	13.33
URM	0.01	0.02	5	84.85	16.97
TF res	0.01	0.07	6	77.42	12.90
TF non-res	0.01	0.01	3	53.33	17.78
SF	0.01	0.01	5	50.00	10.00

Table 1. Determined hyper-parameters (i.e., ν , γ) for the respective SBSTs with affiliated number of SV ($\#SV$), OA estimates (OA [%]) and corresponding values of the evaluation heuristic (OA[%]/ $\#SV$).

A ν -OC-SVM model was trained for each SBST based on the identified hyper-parameter combination and applied to the building inventory. Building objects that were not identified as class members were removed from the *in situ* data. Analogous, instances from the building inventory that were not assigned to one of the 7 SBST classes were removed ([Fig. 3](#)).

3.2.3 Oversampling for Handling Scarce Imbalanced Data Sets

To tackle scarcity of the *in situ* data and address problems associated with class imbalance, we oversampled the training data. As can be seen in [Fig. 2](#), the *in situ* data for the investigated classes exhibit an uneven distribution and for some of them only a very small number of samples is available. If a class is characterized by a small number of samples, it is difficult to uncover regularities and thus construct accurate decision boundaries. In addition, the classifiers are bias-prone and tend to favor the majority classes ([Nguyen et al., 2009](#)).

Accordingly, we employed the SMOTE approach (Synthetic Minority Over-sampling Technique, [Chawla et al., 2002](#)), which generates new samples by interpolating between existing instances rather than simply duplicating original samples to avoid over-fitting. Thereby, the k nearest neighbors of a minority sample within the affiliated minority class are first identified. Depending on the amount of over-sampling required, some of the nearest neighbors are randomly selected and synthetic samples are generated along the line between the minority example and its selected nearest neighbors in the feature space ([He and Garcia, 2009](#)). It has been shown that the error rate caused by class imbalance decreases when the size of the training data increases ([Japkowicz and Stephen, 2002](#)). For this reason, we choose to keep the influence of synthetic samples as low as possible and at the same time equal for each class. Thus, we oversampled *in situ* ground-truth data for all considered classes by 500%, by keeping the *a priori* distribution of the classes. It is ensured that every class consists of at least 60 instances and the classification models deployed later on are not biased due to an alteration of the class distribution. According to the recommendations of [Chawla et al. \(2002\)](#) the number of nearest neighbors to be considered was set to 5.

3.2.4 Multiclass Classification with SVM and RF

For the actual estimation of SBSTs, we deploy SVM and RF since both algorithms showed excellent performances in previous studies, especially when high-dimensional data sets are considered. Contrarily to the OC-SVM approach described in [Section 3.2.2](#), the multiclass SVM approach aims to discriminate two or more target classes from each other. To cope with class overlap or the existence of noise in the training data, soft margin SVM were introduced (referred to as C-SVM). This technique represents a modification of the maximum margin approach using relaxed separation constraints that allow for the possibility of instances on the incorrect side of the respective margin boundary ([Cortes and Vapnik, 1995](#)).

The designated description of SVM refers to binary classification. To extend it to classification problems involving more than two classes normally a parallel architecture made of an ensemble of binary classifiers is utilized ([Hsu and Lin, 2002](#), [Melgani and Bruzzone, 2004](#)). Following this way, one can discriminate one-against-the-rest or one-against-one methods. We chose the latter, because this technique allows producing equal or better performances, while featuring a favorable trade-off between accuracy, computational costs and algorithm complexity ([Hsu and Lin, 2002](#)). As for the OC-SVM approach, also here we used RBF kernels. Learning the most appropriate C-SVM in conjunction with a RBF kernel requires the definition of the cost-parameter C (which controls the trade-off between the maximization of the margin and minimization of the classification error) and the kernel-width parameter γ . Tuning of C and γ was addressed by a grid search strategy based on 10-fold cross validation. Generalization accuracy is evaluated in terms of estimated kappa statistic κ (which allows considering both omission and commission errors; [Foody, 2004](#)) on the average of 5 independent trials. In conformity with the recommendations of [Hsu et al. \(2010\)](#), a coarse grid-search with values of $C = \{2^{-4}, 2^{-3}, \dots, 2^{12}\}$ and $\gamma = \{2^{-5}, 2^{-4}, \dots, 2^3\}$ was performed. Subsequently, a refined grid-search in the neighborhood of the resulting C and γ pair was conducted to determine the final parameterization. A C-SVM was trained for each feature subset ([Section 3.2.1](#)) as well as for the whole feature set.

RF represent a decision-tree-based ensemble learning method for classification and regression introduced by Breiman (2001). Ensemble learning methods build a prediction model by utilizing the strength of a collection of simple base models. Therefore, RF grow multiple decision trees on random subsets of the training data. The high variance among individual trees, letting each tree vote for the class membership, and assigning the respective class according to the majority of the votes, allows the accurate and robust classification of unseen data with little need for fine-tuning, even in the presence of many noisy variables (ibid., Stumpf and Kerle, 2011). Compared to the classification and regression tree approach (Breiman 1984) used in Sarabandi and Kiremidjian (2007), RF represents a bagged predictor and thus always yields higher accuracies (Han and Kamber 2006).

The parameters that need to be specified for generating a RF model consist of the number of classification trees to be grown n_{tree} and the number of features m_{try} used at each node. Both parameters are evaluated by means of the RF inherent out-of-bag (OOB) error measure on the average of 10 independent trials for this study (RF deployed in this paper were carried out with the randomForest package (Liaw and Wiener, 2002) within the statistical computation environment R (R Core Team, 2013)). An increase of n_{tree} causes an improvement of classification accuracy until the performance converges. Since adding more trees to the model does not induce over-fitting, it is possible to run past the point of conversion to obtain more confident OOB values. However, adding more trees simultaneously rises computation time. To determine a suitable tradeoff, models were trained with all features for $n_{tree} = 1, 2, 5, 10, 20, \dots, 50, 100, 200, \dots, 1000, 2000, \dots, 5000, 10000$. We observed that the conversion point for our data set is located between a n_{tree} value of 200 and 300. To provide a reliable OOB error estimate and maintaining the computation times in a reasonable range, we chose a n_{tree} value of 500. This is in a good agreement with the RF parameter study performed by Genuer et al. (2008). According to Breiman (2001) a value for $m_{try} = \sqrt{p}$, with p denoting the number of input features, yields near optimum classification results. However, studies revealed that larger values of m_{try} might perform better for high-dimensional data sets (Breiman, 2002, Genuer et al., 2008). Accordingly, the original training data with all features were used to check for a possible increase in performance. As suggested by Genuer et al. (2008), classification models were learned with nine m_{try} values ($1, \sqrt{p}/2, \sqrt{p}, 2\sqrt{p}, 4\sqrt{p}, p/4, p/2, 3p/4, p$), with $n_{tree} = 500$ for each model. We observed that $m_{try} = \sqrt{p}$ shows near optimum results, coping with the dimensionality of the whole feature set. Thus, we used this parameterization since it is also very favorable from a computational point of view.

4. Results and Discussion

4.1 Detected Outliers

Seven OC-SVM models were learned from the labeled samples to identify and exclude outliers from both the *in situ* data and the building inventory. Each model represents a descriptor of the SBST class it was trained on. Overall, from the 573 labeled samples 32 samples were detected as outliers and thus were removed from the *in situ* data set (Table 2a). Subsequently, each of the seven models was applied separately to the building inventory. The amount of buildings classified as class members and outliers according to each model is shown in Table 2b.

SBSTs	a) <i>in situ</i> data				b) building inventory			
	class members		outliers		class members		outliers	
	abs.	percent	abs.	percent	abs.	percent	abs.	percent
CM	265	97.43	7	2.57	79,543	90.83	8,030	9.17
RC low	186	97.89	4	2.11	77,963	89.03	9,610	10.97
RC high	14	77.78	4	22.22	4,543	5.19	83,030	94.81
URM	29	87.88	4	12.12	63,560	72.58	24,013	27.42
TF res	25	80.65	6	19.35	56,113	64.08	31,460	35.92
TF non-res	12	80.00	3	20.00	29,113	33.24	58,460	66.76
SF	10	71.43	4	28.57	14,188	16.20	73,385	83.80

Table 2 Results of the outlier detection approach for the *in situ* data (a) and building inventory (b) with affiliated numbers and shares of class members and outliers.

As can be seen from Table 2b, notable overlapping between individual classes exists. Especially *CM*, *RC low*, and *URM* show a certain degree of conformity. *TF res* and *TF non-res* seem to be more distinctive from other SBSTs, but feature some interference between each other. *RC high* and *SF* appear to be the most distinctive SBSTs, what seems plausible since these SBSTs are generally characterized by a unique physical appearance. However, the task was not to discriminate one class from another, but to identify instances that feature a distinctive dissimilarity in relation to the *in situ* data. Overall, 5587 instances from the entire building inventory were not assigned to any of the seven SBSTs by the learned models and were therefore excluded.

4.2 Relevance of Features

To identify the most relevant features for building robust multiclass classification models several feature selection algorithms were applied on the data. The features evaluated as most important are shown in Fig. 5. It contains the 15 best ranked features according to the different feature ranking algorithms and the subset revealed from the *CFS* approach. These features contributed the largest shares to raise model performance for both SVM and RF (Section 4.3, Fig. 6).

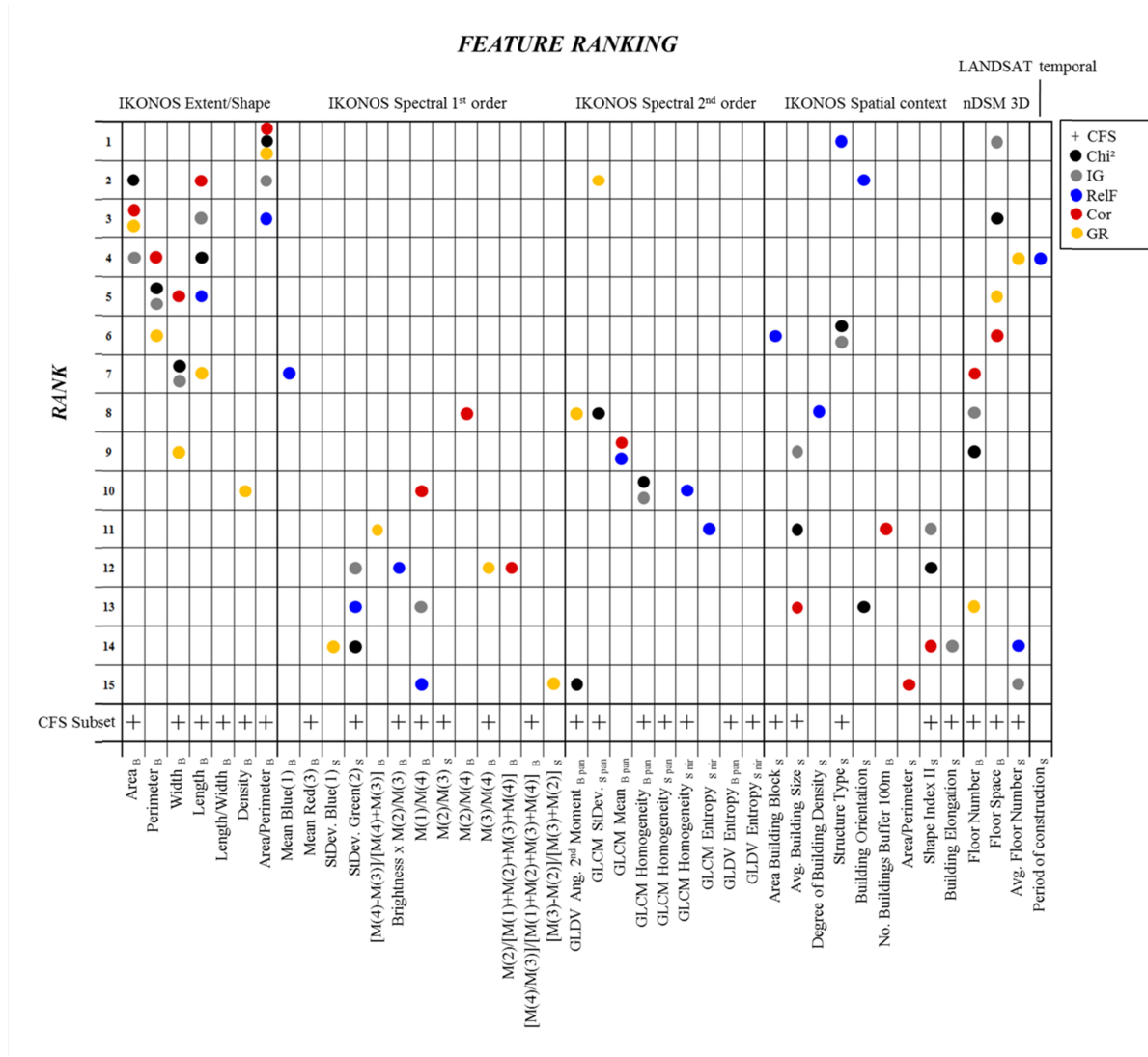


Fig. 5. Benchmarking of the most relevant features. The x-axis lists the features. The y-axis shows the ranks of the respective features according to the feature ranking algorithms and additionally the subset revealed from the *CFS* approach.

Overall, 42 different features of the original 145 feature vector appear in the ranking or the *CFS* subset. 23 features represent individual building characteristics and 19 represent block level information. This shows that both spatial levels are valuable to add specific information for SBSTs estimation. The IKONOS Extent/Shape features are evaluated as very valuable and mostly appear at the top of the ranking. Especially the features “Area/Perimeter_B”, “Length_B”, and “Area_B” have a very prominent position. Our findings hence go in line with earlier work of [Sarabandi and Kiremidjian \(2007\)](#) who stress the importance of features on the extent of buildings to characterize seismic vulnerability. IKONOS Spatial context and nDSM 3D features show a more disperse ranking pattern and thus seem to be slightly less valuable. Interestingly, in relation to [Geiß et al. \(2013\)](#), the expanded feature set ([Fig. 4b](#)) contains some highly ranked features for both categories IKONOS Extent/Shape and IKONOS Spatial context. Features belonging to the groups IKONOS Spectral 1st and 2nd order primarily appear at the bottom positions of the ranking. However, features from these groups are noteworthy

included. Overall, the LANDSAT temporal feature, which is intended to reflect approximate periods of constructions based on data from 1998 and 2000, appears to be less relevant for Padang. However, it is ranked high by the *RelF* approach, which shows distinctive performances for feature sets with a small number of features, as described in the next section.

4.3 Multiclass Classification Models

The classification model performances, decomposed according to the built feature sets, are reported separately for SVM (Fig. 6a) and RF (Fig. 6b). As mentioned, it can be seen that adding the first 15 best ranked features, according to the different feature selection techniques, induces a steep rise of performance. Subsequently, the increase of performance becomes more gradually. When comparing SVM and RF, it can be seen that the performance curve of the RF models start with higher κ values for the subsets containing solely five features. This trend remains until approximately 15-20 features are considered. At this point the model performances become relatively aligned. Adding further features, the estimated κ values of SVM models are higher than the estimated κ values of RF models.

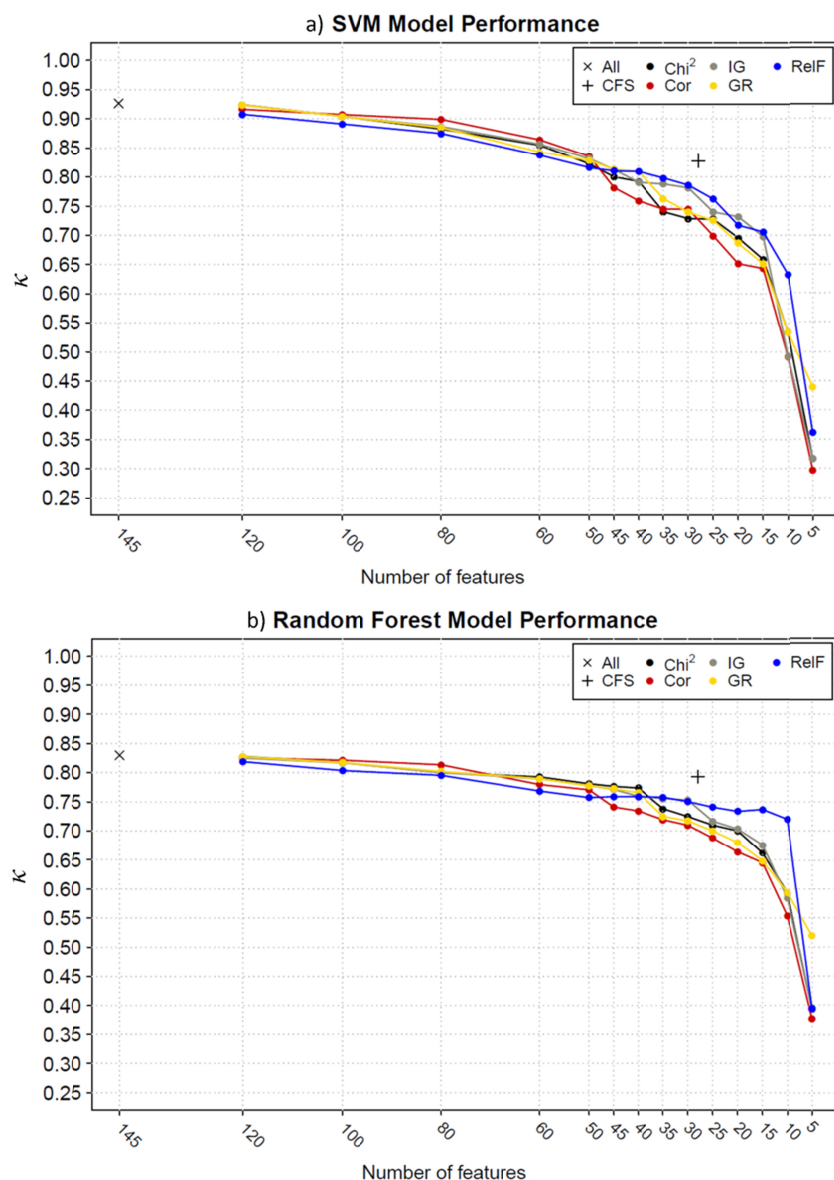


Fig. 6. Performance of the SVM (a) and RF (b) models as a function of the different feature sets. The feature sets were built based on the feature selection techniques described in section 3.2.1. The x-axis lists the number of features contained in the respective feature set and the y-axis reveals the corresponding κ statistic of the learned models.

It is worth mentioning that the RF models built on the features ranked by the *RelF* approach show a comparatively distinctive performance with a few number of features. Especially the ten best ranked features induce a distinctive performance for both SVM and RF, what indicates the usefulness of these features (see Fig. 5) for estimating SBSTs. This feature set is constituted by features from different feature categories. Analogous, the best five features as evaluated by the *GR* approach show distinctive performances and represent also features from different categories. Contrarily, the best five features as evaluated by the *COR* approach belong only to one category and perform worst. Moreover, the *CFS* subset covers a broad spectrum of features from different categories and shows a distinctive positive performance. Overall, this gives an indication that features from different categories are complementary and hence yield higher performances. Regarding the actual κ values, it has to be accentuated that accuracy estimates are very optimistic and do very unlikely reveal real generalization capabilities on unseen data. Due to the application of the SMOTE approach, the training data contains a large fraction of synthetic samples that are very similar to the original *in situ* data. This leads to well bordered clusters in feature space, what makes it possible to learn robust models on the basis of very scarce *in situ* data. Performing a cross-validation with such a data set tends to produce very optimistic accuracy estimates. However, the accuracy estimates in terms of κ statistic served for the comparison and selection of models.

4.4 Classification of Building Inventory and Plausibilization of SBSTs Estimation

For the application and further plausibilization of the approach, we chose the SVM model built with the unreduced feature vector, since it yields the highest estimated κ statistic of all models. The spatially distributed estimation of SBSTs can be seen in Fig. 7a. Analogous to the shares of the different SBSTs of the *in situ* data, the building inventory of Padang is dominated by *CM* and *RC low* buildings. The first is more dominant in remote parts of the city and the latter shapes the central parts. Spatial concentrations of *URM* buildings appear in different sections of the city, whereas *RC high* buildings can be found primarily in the core area parallel to the coast line. *TF* buildings can be found in many parts of the city. They occur both in a diffuse and clustered manner. *SF* buildings appear most dominantly in the south-eastern parts. Outliers primarily represent very huge buildings that are located in central parts or small, informal structures that are located in very remote parts. As described in the previous section, the estimated κ statistic of the models is most likely too optimistic. Thus, we use the *in situ* samples that could not be unambiguously assigned to individual buildings (see Section 2.3) for a further plausibilization. The location of the affiliated building is not exactly known, but must be very close to the respective *in situ* sample.

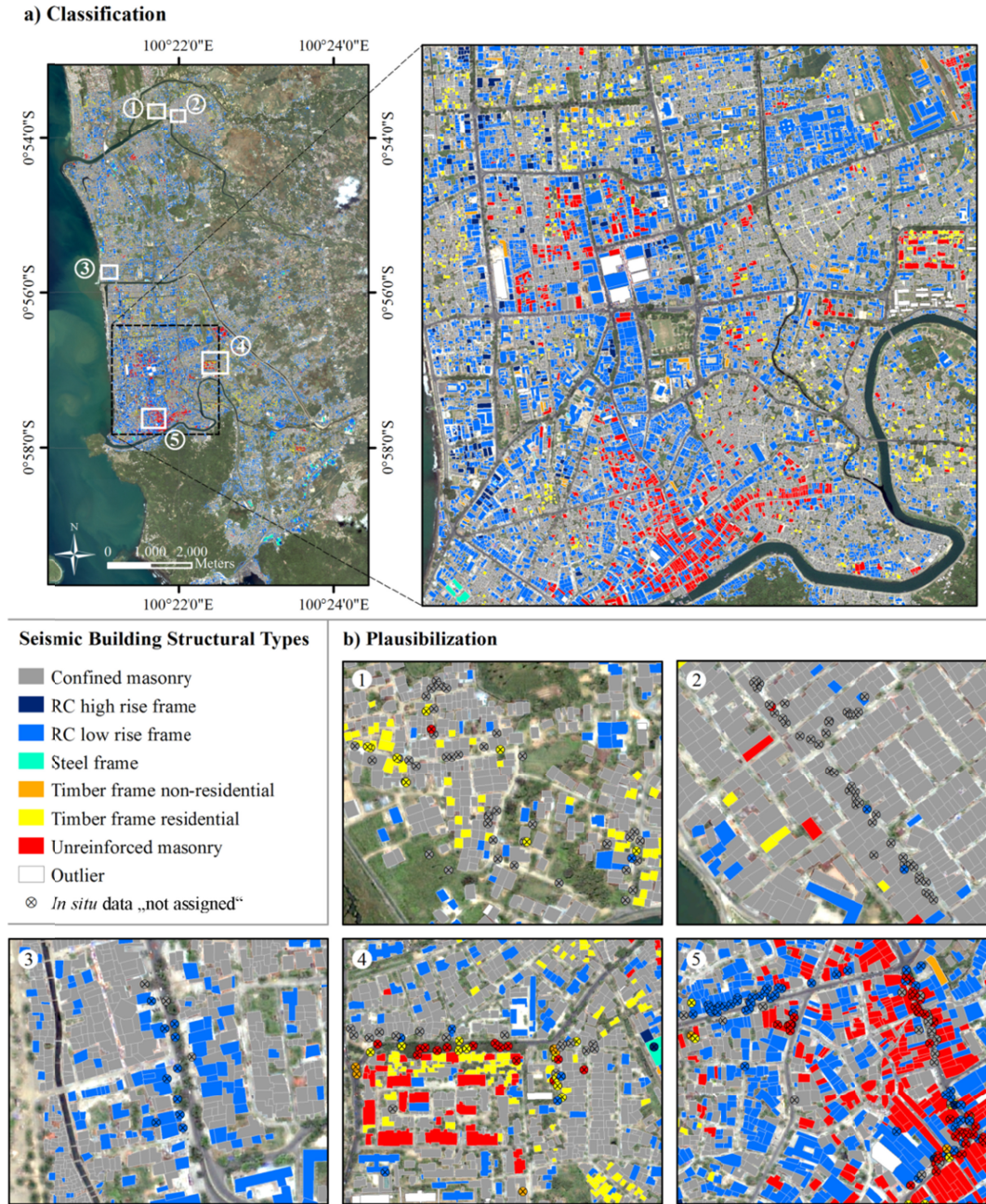


Fig. 7. a) Spatially distributed estimation of SBSTs by application of learned classification model to building inventory; b) Plausibilization of estimation by utilizing the spatially unverified *in situ* samples.

A visual inspection reveals a good agreement in large areas. Five sections of Fig. 7b exemplify the visual plausibility checks. The results appear reasonable especially for the SBSTs *CM*, *RC low*, and *URM*. For timber frame buildings it appears that there were some confusions between *TF res* and *TF non-res*. Due to the fact that the unverified samples do not comprise *SF* buildings and a very small number of *RC high*, we do not make statements about these SBSTs. In order to support validity of estimated accuracies, we checked if there is a corresponding classified building object within a radius of 50 m from each of the 2221

unverified *in situ* buildings. This analysis resulted in an OA of 84%, what underlines the plausibility of the results.

4.5 Application for Earthquake Loss Estimation

For the illustration of the applicability of the approach for ELE modeling, scenario-based loss estimations for Padang are presented. Fragility functions in the form of cumulative lognormal distributions have been derived from data collected after the 30th September 2009 event (Fig. 8a, Sengara et al., 2010) for different SBSTs. They relate the Modified Mercalli Intensity (MMI) to a damage index *DI*. The latter represents an economic measure for damage and is constituted by the ratio of repair cost and total building reconstruction cost. For the presented scenarios, we assume that site effects do not play a significant role in the intensity distribution. Thus, the MMI is assumed to be spatially invariant within the city of Padang, what is consistent with observations from the aforementioned event (*ibid.*). Accordingly, Figure 8b reveals calculated building inventory loss for several MMIs for the presented study area. In addition, spatially distributed building damage based on a *violent* MMI of 9 is shown in Fig. 8c. The estimation of densities regarding damaged buildings was carried out with a *KDE* approach (Bailey and Gatrell, 1995), with a Gaussian kernel and an adaptive kernel bandwidth that ensures a sample size of 50 buildings. Note that “*SF*” buildings and outliers are not included in the analyses due to missing fragility functions.

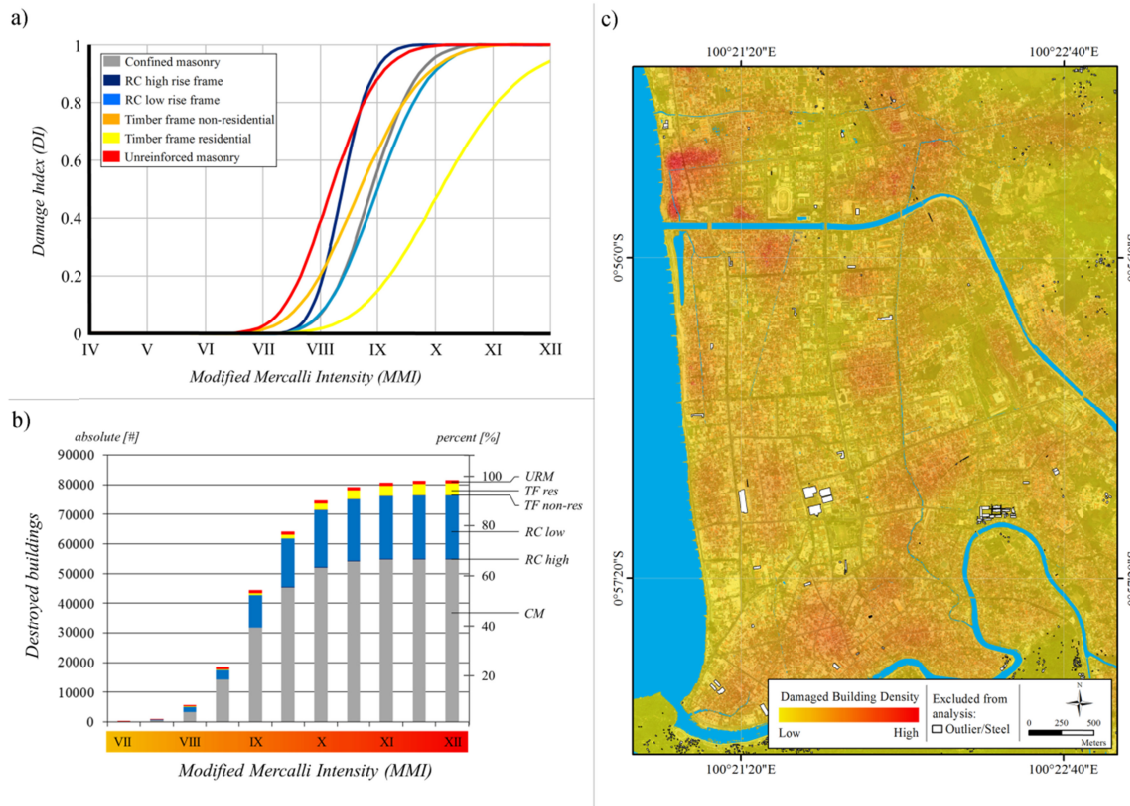


Fig. 8. Application for scenario-based ELE; a) Fragility functions for different SBSTs. They were derived empirically by Sengara et al. (2010) for the buildings of Padang after the 30th

September 2009 earthquake; b) building inventory loss for several MMIs; c) spatially distributed building damage for a MMI of 9.

The presented approach allows to quantify building damage in a detailed way and make damage estimations spatially explicit by e.g., localizing hot spots within a city. These are key features for earthquake loss modeling and predictions.

5. Conclusions and Outlook

This study aims to demonstrate how to estimate seismic building structural types by means of a remote sensing based approach and to evaluate the suitability of the data and methods. As there is no obvious correlation between the digital information of pixels and SBSTs, we showed how to derive sets of valuable features to characterize the urban environment. For this purpose we utilized very high resolution multispectral EO-data, height information and multi-temporal medium resolution multispectral EO-data. It turned out that a combinatory use of features from different feature categories related to building shape and height, spatial context, and spectral information appears most promising. In this manner, we proposed a hierarchical supervised classification scheme that adapts techniques from the machine learning domain to estimate SBSTs. By means of a sequential procedure including *feature selection*, *outlier detection*, *generation of synthetic samples*, and *learning non-parametric SVM and RF classification models*, SBSTs could be estimated spatially distributed with plausible accuracies. Thus, we conclude that remote sensing data and methods have a high capability to support large area estimation of SBSTs.

A spatially distributed and accurate estimation of SBSTs is a critical input for ELE models as illustrated in [Section 4.5](#) and thus represents a relevant contribution to seismic risk mitigation and preparedness activities. Furthermore, related to an actual earthquake hazard, post-event earthquake loss estimations for response and recovery activities can be made, e.g. also advanced earthquake early warning systems include the capacity for the rapid assessment of damage ([Picozzi et al., 2013](#)). Such information would also be useful for a guided and thus more accurate post-event damage mapping ([Dell’Acqua and Gamba, 2012](#)). Beyond, remote sensing has the capability to quantify exposed people ([Taubenböck et al., 2009b](#)), what also allows the estimation of human casualties within the aforementioned model setting.

However, from a technical perspective, future research can utilize the enhanced resolution characteristics of latest and future spaceborne missions, such as WorldView-2 and 3, CARTOSAT-3 or ALOS-3, what allows the calculation of more sophisticated feature vectors ([Novack et al., 2011](#)). In addition, the application of computationally intensive wrapper methods for feature selection should be investigated. Regarding the actual classification task, the application of a semisupervised approach (e.g. [Bruzzone et al., 2006](#)), which also encodes some knowledge from the unlabeled data, appears promising especially when only very few labeled samples are available. To this end, active learning methods (e.g. [Tuia et al., 2009](#)) enable a guided selection of the most feasible samples. Nevertheless, the most exigent task regarding future research comprises the systematic and comprehensive collection of accurate georeferenced *in situ* data for both SBSTs and experienced earthquake damage by structural

engineers. Only this way the remote sensing community can fully demonstrate the usability of EO-data for SBSTs estimation and ELE modeling. This is why we see a need to trigger an open dialogue between the remote sensing and earthquake engineering community to share data and gather a common understanding about e.g., typologies. We believe that only a close interdisciplinary collaboration will enable systematic and valid large-area estimations of SBSTs and earthquake loss of dynamic earthquake prone urban areas around the globe.

Acknowledgments

The research was founded partially by Helmholtz-EOS (Earth Observation System). The authors also would like to thank the DFG/BMBF special Programme “Geotechnologies”—Early Warning Systems in Earth Management. Sponsorship Code: 03G0643A-E. Furthermore, we highly appreciate the encouragement of Jonathan Griffin (Geoscience Australia) and the Australia-Indonesia Facility for Disaster Reduction (AIFDR) regarding the cooperation and data sharing.

References

- Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis, John Wiley and Sons, New York.
- BNPB – Badan Nasional Penanggulangan Bencana (National Disaster Management Agency, Indonesia), 2009. Laporan Harian Pusdalops BNPB Minggu, <http://www.bnpb.go.id> . (Accessed 18 May, 2013)
- Borfecchia, F., Pollino, M., De Cecco, L., Lugari, A., Martini, S., La Porta, L., Ristatore, E., Pascale, C., 2009. Active and passive remote sensing for supporting the evaluation of the urban seismic vulnerability. *Italian Journal of Remote Sensing* 42, 129–141.
- Borzi, B., Dell'Acqua, F., Faravelli, M., Gamba, P., Lisini, G., Onida, M., Polli, D., 2011. Vulnerability study on a large industrial area using satellite remotely sensed images. *B. Earthq. Eng.* 9, 675–690.
- Breiman, L., 2001. Random Forests. *Mach. Learn.*, 45, 5–32.
- Breiman, L., 2002. Manual on setting up, using, and understanding random forests V3.1. Department of Statistics, University of California, Berkeley, http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf . (Accessed 8 Feb, 2013)
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth.
- Bruzzzone, L., Chi, M., Marconcini, M., 2006. A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 44 (11), 3363–3373.

Calvi, G.M., Pinho, R., Magenes, G., Bommer, J.J., Restrepo-Vélez, L.F., Crowley, H., 2006. Development of seismic vulnerability assessment methodologies over the past 30 years. *ISET Journal of Earthquake Technology* 43 (3), 75–104.

Camps-Valls, G., Bruzzone, L., 2009. *Kernel Methods for Remote Sensing Data Analysis*, John Wiley & Sons, New York.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* 16, 321–357.

Chang, C.C., Lin, C.J., 2001. LIBSVM: A library for support vector machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm> . (Accessed 12 Jan, 2013)

Chlieh, M., Avouac, J.P., Sieh, K., Natawidjaja, D.H., Galetzka, J., 2008. Heterogeneous coupling of the sumatran megathrust constrained by geodetic and paleogeodetic measurements. *J. Geophys. Res.* 113 (B5), B05305.

Coburn, A., Spence, R., 2002. *Earthquake Protection*, John Wiley & Sons, New York.

Colaninno, N., Roca, J., Pfeffer, K., 2011. An automatic classification of urban texture: form and compactness of morphological homogeneous structures in Barcelona. 51st European Congress of the Regional Science Association International, pp. 1–20.

Cortes, C., Vapnik, V., 1995. Support vector networks. *Mach. Learn.* 20 (3), 273–297.

Dell'Acqua, F., Gamba, P., 2012. Remote sensing and earthquake damage assessment: Experiences, limits, and perspectives. *P. IEEE* 100 (10), 2876–2890.

Douglas, J., 2007. Physical vulnerability modelling in natural hazard risk assessment. *Nat. Hazard. Earth Sys.* 7 (2), 283–288.

Duch, W., 2006. Filter methods, Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing), Guyon, I., Gunn, S., Nikraves, M., Zadeh, L.A., Eds., chapter 3, pp. 90–117, Springer, New York.

Dunbar, P.K., Bilham, R.G., Laituri, M.J., 2003. Earthquake Loss Estimation for India Based on Macroeconomic Indicators, *Risk Science and Sustainability*, NATO Science 112, 163–180.

EERI – Earthquake Engineering Research Institute, 2009. Learning from Earthquakes – The M_w 7.6 Western Sumatra Earthquake of September 30, 2009, EERI Special Earthquake Report, Dec. 2009, USA.

Fayyad, U.M., Irani, K.B., 1993. Multi-interval discretization of continuous-valued attributes for classification learning, in *IJCAI*, pp. 1022–1029.

Foody, G.M., 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogramm. Eng. Remote Sens.* 70, 627–633.

Geiß, C., Taubenböck, H., 2012. Remote sensing contributing to assess earthquake risk: from a literature review towards a roadmap. *Nat. Hazards*, 68, 7–48.

Geiß, C., Taubenböck, H., Tyagunov, S., Tisch, A., Post, J., Lakes, T., 2013. Assessment of seismic building vulnerability from space. *Earthq. Spectra*, in press, doi: 10.1193/121812EQS350M.

Genuer, R., Poggi, J.M., Tuleau, C., 2008. Random forests: some methodological insights, Rapport de recherche RR-6729, INRIA.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. *Pattern Recogn. Lett.* 27, 294–300.

Gravelius, H., 1914. Grundriß der gesamten Gewässerkunde. Band I: Flußkunde (Compendium of Hydrology, vol. I. Rivers, in German), Goschen, Berlin, Germany.

Guyon, I., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Grünthal, G., Musson, R., Schwarz, J., and Stucchi, M., 1998. European Macroseismic Scale, Cahiers de Centre Européen de Géodynamique et de Seismologie, vol 15, Luxembourg.

Hall, M.A., 1999. Correlation-based Feature Selection for Machine Learning, Ph.D. dissertation, Department of Computer Science, The University of Waikato, Hamilton, New Zealand.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11 (1), 10–18.

Hall, M., Holmes, G., 2003. Benchmarking attribute selection techniques for discrete class datamining. *IEEE T. Knowl. Data En.* 15 (6), 1437–1447.

Han, J. and Kamber, M., 2006. Data mining: concepts and techniques. Elsevier, Morgan Kaufmann, 2nd edition.

He, H., Garcia, E., 2009. Learning from imbalanced data. *IEEE T. Knowl. Data En.* 21 (9), 1263–1284.

Herold, M., Liu, X., Clarke, K., 2003. Spatial metrics and image texture for mapping urban land use, *Photogramm. Eng. Rem. S.* 69, 991–1001.

Hsu, C.W., Chang, C.C., Lin, C.J., 2010. A practical guide to support vector classification, Technical report, Department of Computer Science, National Taiwan University.

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multi-class support vector machines. *IEEE T. Neural Networ.* 13(2), 415–425.

Hughes, G.F., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE T. Inform. Theory* 14 (1), 55–63.

Intermap, 2010. Product handbook and quick start guide. Standard edition, v4.4, <http://www.intermap.com/images/handbook/producthandbook.pdf> . (Accessed 7 Jul, 2012)

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. *Intell. Data Anal.* 6 (5), 429–449.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. *Artif. Intell.* 97 (1), 273–324.

Kononenko, I, 1994. Estimating attributes: Analysis and extensions of relief, in *European Conference on Machine Learning*, Bergadano, F., L.D. Raedt, Eds., Springer, pp. 171–182.

Li, X., Tennant, K., Lawrence, G., 2004. Three-dimensional mapping with airborne IFSAR based STAR technology – Intermap’s experiences, *International Archives Photogrammetry and Remote Sensing Vol. XXXV-B3*, pp. 261–266.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News: The Newsletter of the R Project* 2 (3), 18–22.

MacEachren, A.M., 1985. Compactness of geographic shape: comparison and evaluation of measures. *Geogr. Ann. B.* 67 (1), 53–67.

McCaffrey, R., 2009. The Tectonic Framework of the Sumatran Subduction Zone. *Annu. Rev. Earth Pl. Sc.* 37, 345–366.

McCloskey, J., Lange, D., Tilmann, F., Nalbant, S.S., Bell, A.F., Natawidjaja, D.H., Rietbrock, A., 2010. The September 2009 Padang earthquake. *Nat. Geosci.* 3, 70–71.

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* 42 (8), 1778–1790.

Muñoz-Marí, J., Bovolo, F., Gomez-Chova, L., Bruzzone, L., Camp-Valls, G., 2010. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Trans. Geosci. Remote Sens.* 48 (8), 3188–3197.

Nguyen, G.H., Bouzerdoun, A., Phung, A., 2008. A supervised learning approach for imbalanced data sets. *Pattern Recogn.*, 1–4.

Novack, T., Esch, T., Kux, H., Stilla, U., 2011. Machine learning comparison between worldview-2 and quickbird-2-simulated imagery regarding object-based urban land cover classification. *Remote Sensing* 3 (10), 2263–2282.

Petersen, M.D., Dewey, J., Hartzell, S., Mueller, C., Harmsen, S., Frankel, A.D., Rukstales, K., 2004. Probabilistic seismic hazard analysis for Sumatra, Indonesia and across the Southern Malaysian Peninsula. *Tectonophysics* 390 (1-4), 141-158.

Picozzi, M., Bindi, D., Pittore, M., Kieling, K., Parolai, S., 2013. Real-time risk assessment in seismic early warning and rapid response: a feasibility study in Bishkek (Kyrgyzstan). *J Seismol* 17, 485-505.

Pittore M., Wieland, M., 2012. Towards a rapid probabilistic seismic vulnerability assessment using satellite and ground-based remote sensing. *Nat. Hazards*, doi: 10.1007/s11069-012-0475-z.

R Core Team, 2013. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

Richter, R., 1996. A spatially adaptive fast atmospheric correction algorithm. *Int. J. Remote Sens.* 17, 1201–1214.

Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.* 53 (1-2), 23-69.

Sarabandi, P., Kiremidjian, A., 2007. Development of algorithms for building inventory compilation through remote sensing and statistical inferencing, The John A. Blume Earthquake Engineering Center, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA, Report No. 158, 419p.

Schölkopf, B., Smola, A., 2002. *Learning with Kernels*. MIT Press, Cambridge (MA).

Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., 1999. Support vector method for novelty detection, *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, vol. 12, pp. 582–588.

Sengara, W.I., Suarjana, M., Beetham, D., Corby, N., Edwards, M., Griffith, M., Wehner, M., Weller, R., 2010. The 30th September 2009 West Sumatra Earthquake: Padang region damage survey, Geoscience Australia, Record 2010/44, 201p.

Steiniger, S., Lange, T., Burghardt, D., Weibel, R., 2008. An approach for the classification of urban building structures based on discriminant analysis techniques. *Transactions in GIS* 12 (1), 31–59.

Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using random forests. *Remote Sens. Environ.* 115 (10), 2564–2577.

Taubenböck, H., Goseberg, N., Lämmel, G., Setiadi, N., Schlurmann, T., Nagel, K., Siegert, F., Birkmann, J., Traub, K.-P., Dech, S., Keuck, V., Lehmann, F., Strunz, G., Klüpfel, H., 2013. Risk reduction at the „Last-Mile“: an attempt to turn science into action by the example of Padang, Indonesia. *Nat. Hazards* (65), 915-945.

Taubenböck, H., Post, J., Kiefl, R., Roth, A., Ismail, F.A., Strunz, G., Dech, S., 2008. Risk and vulnerability assessment to tsunami hazard using very high resolution satellite data, in *Proceedings of the EARSel Joint Workshop: Remote Sensing: New Challenges of high resolution*, J. Carsten, Ed., Bochum, Germany.

Taubenböck, H., Roth, A., Dech, S., Mehl, H., Munich, J.C., Stempniewski, L., Zschau, J., 2009a. Assessing building vulnerability using synergistically remote sensing and civil engineering, in *Urban and Regional Data Management*, Kreck, A., Rumor, M., Zlatanova, S., Fendel, E., Eds., Taylor & Francis Group, London, pp. 287–300.

Taubenböck, H., Goseberg, N., Setiadi, N., Lämmel, G., Moder, F., Oczipka, M., Klüpfel, H., Wahl, R., Schlurmann, T., Strunz, G., Birkmann, J., Nagel, K., Siegert, F., Lehmann, F., Dech, S., Gress, A., Klein, R., 2009b. « Last-Mile » preparation for a potential disaster – Interdisciplinary approach towards tsunami early warning and an evacuation information system for the coastal city of Padang, Indonesia. *Nat. Hazard. Earth Sys.* 9, 1509–1528.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M.F., Emery, W.J., 2009. Active Learning Methods for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* 47 (7), 2218–2232, 2009.

Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., Wald, R., 2009. Feature selection with high-dimensional imbalanced data, in Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, ICDMW '09, pp. 507–514, Washington, DC, USA. IEEE Computer Society.

Vapnik, V.N., 1995. The nature of statistical learning theory, New York, Springer.

Vapnik, V.N., 1998. Statistical learning theory. New York, Wiley.

Volpi, M., Tuia, D., Bovolo, F., Kanevski, M., Bruzzone, L., 2013. Supervised change detection in VHR images using contextual information and support vector machines. *Int. J. Appl. Earth Obs.* 20, 77–85.

Wieland, M., Pittore, M., Parolai, S., Zschau, J., Moldobekov, B., Begaliev, U., 2012. Estimating building inventory for rapid seismic vulnerability assessment: towards an integrated approach based on multisource imaging. *Soil. Dyn. Earthq. Eng.* 36, 70–83.